Volume 5, Number 7
July 1991

Long, Fat Pipes

by Carl Malamud

[Ed.: The following article is adapted from the book *Stacks—Interoperability in Today's Computer Networks.* The book will be published at INTEROP 91 in October].

Introduction

"But why would I want a gigabit to my desktop?" I had been enthusiastically describing the Gigabit Testbed projects coordinated by the Corporation for National Research Initiatives to a prominent member of the industry. Gigabits of network bandwidth, petabytes of secondary storage, and teraflop computers had all been bandied about in a description of tomorrow's high-speed networks.

"Why would I want a gigabit?" is similar to a question that was common a few years ago, "Why would a PC ever need a full 640 kbytes of memory?" Needless to say, as soon as people discovered spreadsheets, 640 kbytes was not only reasonable, it became the minimal acceptable amount. When we learn to make do with what we have, we sometimes forget that the driving force is not our ability to make do with the existing technological base but the demand by users to get work done.

To see why we might want gigabit networks, let's start again with the lowly PC. If you want to do computer-generated real-time graphics, think of the VGA interface on the PC. A VGA screen has 640 x 480 bits with 256 simultaneous colors. To support 256 simultaneous colors, you need one byte per pixel. If you are operating at 30 screens per second, generally recognized as the minimum acceptable rate for real-time video, you are generating a data rate of 73.728 Mbps.

Now extend this analysis to workstations. Consider screens with 1000 x 1000 pixels and at least 2 bytes per pixel (yielding 64,000 simultaneous colors). All of a sudden we have increased our data rate from 73.728 Mbps to 480 Mbps.

Not every user needs 480 megabits per second on a transcontinental basis. Not every user is going to need all this bandwidth at all times. However, a few users will need this bandwidth at a time, and there are many, many users on real networks. We need both the capacity to deliver this bandwidth to the individual as well as the aggregate bandwidth to handle large numbers of users.

Another way to see the demand for high-speed networks is to examine how other portions of the computing environment are growing. A high-level (but not unusual) personal workstation has 100 Mbytes of storage, operates at 1–20 MIPS, and has 4–8 Mbytes of main memory. A rule that has held true for many years is that the shared computer of today ends up being the personal computer of tomorrow. A 1–20 MIPS machine with 8 Megabytes of Memory a few years ago would have been a large, shared VAX, but is now a personal workstation.

To see what the personal computer of tomorrow will be like, look at today's larger systems. It is not at all unusual to see 1 Gbyte of secondary storage, 20–50 MIPS systems, and 32–128 Mbytes of main memory. In fact, it is now possible (though fairly expensive) to buy personal workstations in this range. Over time, workstations will start to reach these levels for large numbers of people.

Let us look at even larger systems. Supercomputer centers and research laboratories are already working with terabytes of data. Groups like NASA are beginning to think in terms of petabytes (thousands of terabytes) of secondary and tertiary storage and some people are beginning to think in terms of exabytes (millions of terabytes).

Current large scale processors operate in the billion operations per second range. The *High Performance Computing and Communications* (HPCC) initiative will pour serious money into the development of a computer that will operate at a trillion operations per second.

Finally, look at main memory. Large supercomputer installations like the NASA-Ames Research Center have systems with main memories in the gigabyte range. In addition to a gigabyte of main memory, these systems often have another gigabyte or two allocated as a RAM disk. Systems with a terabyte or more of main memory are not that far off.

Balance is the key to any computer configuration. If the machines are faster and the disk drives larger, the networks also need to grow. If you need to load data into a terabyte of main memory, you are going to need more than an Ethernet.

The rationale for high-speed networks is particularly compelling if you realize that certain computer facilities will not be able to be duplicated. Computers are expensive, particularly supercomputers. In many cases, it won't make sense to buy one of each for every site. Instead, we need to put different computers in different locations.

Users will need to put these disparate computing locations together to form solutions to problems. Many efforts are now underway that examine how different computing environments can be joined together to solve specific problems. In this article, we will look at two of these efforts.

Both the networks examined in this article, *VISTAnet* and *CASA*, are part of the National Gigabit Testbed program, coordinated by CNRI, the Corporation for National Research Initiatives. There are three other testbeds as part of the project. Funding for this program of $15.8 million comes from DARPA and NSF, with another $100 million in facilities, equipment, and personnel thrown in by a list of industrial participants that includes almost every major computer and telecommunications company in the country.

VISTAnet

The VISTAnet project is coordinated by the MCNC, a non-profit corporation that runs the North Carolina Supercomputing Center. The group also runs another network called CONCERT which is a statewide private network built strictly on microwave towers. The network consists of three video networks that deliver NTSC signals to classrooms for classes and teleconferencing, plus a data network.

Computers

The VISTAnet project brings together three types of computers. First (of course) there is a large *Cray* computer. In this case the Cray computer is a CRAY Y-MP 8/432 with four processors, 64 megawords of main memory and another 128 megawords in a solid state disk. The machine has a peak performance of 1.2 Gflops. The Cray computer is located in a supercomputer center in the Research Triangle Park in North Carolina.

The second computer is a *Pixel-Planes 5*, an experimental machine developed at the Computer Science Department at the University of North Carolina. This machine does high-speed rendering of graphics. This autistic computer has the ability to do this one type of operation, and only this one type of operation, very fast. The Pixel-Planes is ideal for rendering polygonal images with lighting, shadows, and textures. The third machine is a *MasPar*, a commercial parallel processor, used for performing statistical manipulations.

continued on next page

Long, Fat Pipes *(continued)*

All three of these machines are combined to help feed a workstation used by the Department of Radiation Oncology at UNC. The machine uses a joy stick to allow the physician to control a 1280 x 512 pixel color display that shows a three-dimensional representation of radiation doses.

The VISTAnet project links machines in three locations: the North Carolina Supercomputing Center (NCSC) in Research Triangle Park, and two departments at the University of North Carolina, the Department of Radiation Oncology and the Computer Sciences Department.

Switch technology

The three customer locations are near two different central offices, each operated by a different telephone company. University of North Carolina links up to a Southern Bell office. Research Triangle Park is served by GTE.

The two central offices are linked together with a 2.4 Gbps OC-48 SONET line. A Fujitsu FETEX-150B-ISDN ATM switch is placed in the Southern Bell office and provides OC-12C links to the two UNC departments, each link operating at 622 Mbps.

The Fujitsu switch is the primary switch for the network. The link to the OC-48 line moves data over to the GTE office. There, a broadband circuit switch moves data on toward the Cray computer. The circuit switch (also known as a digital cross connect) can also be used for other applications such as video teleconferencing.

Because computers do not have a raw SONET link, another standard is used to move the data onto the computer systems. A HIPPI to ATM *Network Terminal Adaptor* (NTA) provides this function. The computers have a simple HIPPI interface to the NTA.

The role of the NTA is to take incoming ATM cells and present them to the HIPPI interface at 800 Mbps. The NTA has to perform rate adaption between the ATM rate of 622 Mbps and HIPPI's 800 Mbps. The NTA also provides the connection management function. Remember that the HIPPI interface allows communication with one device at a time, even though the ATM interface allows multiplexing of traffic. The NTA blocks calls until a virtual circuit is available to a remote HIPPI interface.

Why VISTAnet?

VISTAnet is in place mainly to do networking research. It sure helps, however, when you have a user. The user for VISTAnet is a fascinating experiment that applies high-speed networking to an area of medical practice known as *radiation oncology.*

When a person gets cancer, there are three ways to treat it. Surgery and chemotherapy are often used, but suffer from many drawbacks. A third approach is to use *radiation therapy.* Radiation therapy takes a cancer and kills it with a beam of radiation. The problem is that both normal and diseased cells get killed when exposed to radiation. Since the beam must pass through healthy tissue, a beam strong enough to kill the cancer will also kill healthy cells.

Luckily, the effect of radiation on cells is dependent on the dose. Small exposures do not hurt cells. If we split a radiation dose up into several beams that intersect at the diseased area, we can kill the diseased cells because they receive exposure to all beams. Healthy cells only receive exposure to a single beam and thus are able to survive.

Planning radiation treatment strategies starts with a CT scan of the diseased and surrounding areas. Because each cancer is unique—each has its own location, shape, and size—a doctor must develop an individual treatment plan for each situation that kills the diseased area without killing healthy cells.

Developing treatment plans operates in a very large parameter space with an infinite number of possible solutions. CT scans allow a two-dimensional view of the area. Analysis of a treatment plan shows the distribution of the dose over diseased and healthy areas within a single plane, but does not show the effect of the beams above and below the plane.

Problems

We thus have two problems. First, the number of possible solutions is very large. The number of beams, the locations of the beams, the position of the patient, and the type of shielding are just a few of the parameters. The complexity of the problem means only a few possible plans are tried.

The second problem is that the two-dimensional nature of the analysis means that only a portion of the effect can be seen. The result is that it is not uncommon for a treatment to get most of a diseased area, but not all, known as a *local failure.*

Estimates are that over 390,000 patients are treated per year with radiation therapy. Of those, 38,000 of the treatments are subject to local failures. While better diagnosis and treatment plans would not save all of those patients, it is likely that several thousand lives could be saved with better treatment.

Radiation oncology is thus a great application for high speed networking. When the VISTAnet project was trying to find an application for a gigabit testbed, they approached Dr. Julian Rosenman at the University of North Carolina.

Rosenman explained his problem in radiation oncology and his large computational requirements. To calculate a radiation dose distribution, a model consists of a system of 256^3 points. Each of these data points occupies 16 bits per point, resulting in a data rate of 256 megabits, clearly the province of a Cray computer.

Graphic representation

The information coming out at this rate is not very useful as raw data. It needs to be graphically represented to be useful to a physician looking at alternative treatment plans. The data stream goes into the Pixel-Planes 5 machine, 18 miles away from the Cray computer. This system is able to render the incoming data stream in near real time, at which point it needs to be displayed on a workstation, resulting in another very large data stream going over to the workstation.

VISTAnet is an example of how a single user can easily use a gigabit network. Visualization of radiation doses is an application that could not work in one site. It requires scarce facilities in multiple locations.

Note that in the medical field, it is not just the supercomputers that are scarce resources. Medical equipment is often very expensive and cannot be duplicated. Networks allow this scarce medical equipment to be used along with other scarce resources in other locations to form a more complete picture of diagnosis or treatment.

continued on next page

Long, Fat Pipes *(continued)*

CASA

A second gigabit testbed project is CASA. CASA involves four of the most highly developed computing centers in the country:

- San Diego Supercomputer Center (SDSC)
- California Institute of Technology (Caltech)
- Jet Propulsion Laboratory (JPL)
- Los Alamos National Laboratory (LANL)

All four of these sites are known for providing the latest in supercomputer facilities. Caltech and LANL are both leaders in applying parallel processors such as the Connection Machine to real-world problems.

With all these large facilities, however, there are a series of problems that can overwhelm any one of the computer centers. CASA will tie all four of the sites together with an 800 Mbps computer network, spanning up to 1300 kilometers.

CASA involves three applications, two of which are described in this article, that require very high-speed networks. Like the VISTAnet radiation oncology example, they are real-world problems that require solutions unavailable on any one large computer system, or even in any one large computer center.

Predicting the weather

Weather modeling is one of the applications that helps to spur larger and larger computers. Our weather system is so complex that most models concentrate on either the ocean or the atmosphere. Even dividing the problem into two leads to immense amounts of data.

Take atmospheric models, for example. If we take the world and divide it into "squares" of five degrees longitude by four degrees latitude, we have a fairly coarse grid of the world. If we model nine altitude layers, we have a grid of 72 x 44 x 9.

Even this coarse model of the atmosphere requires ten CPU seconds on a CRAY X-MP/48 to advance the model one hour. If we want to study a particular weather phenomenon, such as the Greenhouse Effect, it is not unusual to run a model through 50 years of space, requiring about 35 CPU days on the Cray computer.

Remember, this simplified model represents only one-half of the weather system, the atmosphere. The ocean model, at a coarse approximation, is a grid of 360 x 180, 27 levels deep. To advance this model a single hour, takes 20 CPU seconds on the CRAY X-MP/48, twice as long as the atmospheric model.

Although the atmosphere and the ocean have been separated, the weather should really be treated as a closed system. The output from the ocean model, especially sea surface temperature, is a key input to the atmospheric model. Outputs from the atmospheric model, such as winds and heat flux, are key inputs to the ocean model.

Under the direction of R. Mechoso of UCLA, CASA will combine two standard models to form a single closed system, the input from one model driving the other. Two machines will be used, one for each model.

The oceanic model will be put on a Connection Machines CM-2, which speeds the ocean model up by a factor of 50–100 times. The speedup is due to the superior programming model, for this particular application, of the massively parallel architecture.

The atmospheric model will be put on a CRAY Y-MP 8/864, located at the San Diego Supercomputer Center. This particular configuration of the CRAY Y-MP yields speedup of twelve times over the CRAY X-MP. Notice that the ocean model will be running fifty times faster, whereas the atmospheric model will run only a dozen times faster.

To handle the mismatch, the hydrodynamic part of the atmospheric model will be moved over to the CM-2, leaving the Cray computer with atmospheric problems like cumulus cloud convection and radiation calculations. Of course, a tremendous amount of data needs to move between the Cray computer and the CM-2, including data such as temperature and humidity for each grid for each cycle. Estimates are that approximately 750 Mbps per second of data will be transferred between the two machines.

Why bother with all this? A unified model is a way of tuning individual components so they reflect reality much more closely. If valid inputs yield valid outputs, we can start looking more accurately at questions like the Greenhouse Effect, forecasts of trade winds, and other global phenomena.

3-D Seismic profiling

Before we look at the CASA network itself, we examine one more CASA application, involving three-dimensional rendering of data from multiple earth-science data sets. This project is run at the Jet Propulsion Laboratory, but takes advantage of data and computers in different CASA locations.

Data in the earth sciences is increasing at fairly astonishing rates. There are a variety of different sources of this information: LANDSAT, topographic databases, and seismic databases, for example. One of the real challenging sources of information will be the space station, known as the NASA *Earth Orbiting System* (EOS). The EOS will be sending data down at the rate of 300 Mbps, equivalent to ten Gbytes every six minutes. And this is just one of many sources.

Combining information from different sources allows a variety of very important applications, including the modeling of earthquake faults, which allows prediction of an estimate of the order of magnitude of a coming earthquake (but not the exact time).

Earth sciences databases can be used for a variety of other tasks. Combined data sets have allowed researchers to discover that the Sahara desert was once a large river basin and even to find long-hidden roadways in Mongolia and Arabia, buried for several thousand years.

The point of the CASA application is to try and learn how to handle these very large datasets coming from different locations. For the JPL, this project is preparation for the flood of data expected from the space station. JPL is trying to learn how to handle data streams of three Gbytes/second and up which could require 90 Gigaflops or more to process.

Being able to handle data quickly is often crucial. An example is when the Voyager-2 was approaching Neptune. When the Voyager was 3–4 days out from the closest approach, an interesting feature was found on Neptune. Normally, it would take VAX systems weeks to analyze the data and provide positioning instructions for the on-board cameras. Instead, an eight-node Mark IIIfp was used to make the calculation quickly enough to send up repositioning instructions.

continued on next page

Long, Fat Pipes *(continued)*

The particular application chosen will merge data from three sources to provide 3-D cutaways of the earth's surface, allowing the identification of fault zones and major plate thrusts. Interactive 3-D graphics are essential for this application, because researchers cannot tell ahead of time the level of detail and particular view they need when examining specific places in the earth.

The three sources of information include the LANDSAT thematic mapper, CALCRUST seismic reflection data, and elevation data from the Space Shuttle's imaging radar. The amount of data involved for each image produced is fairly amazing.

Filtering

The LANDSAT thematic mapper, for example, involves a typical image of 90 x 90 kilometers. The image is broken up into 3000 by 3000 pixels with seven bands at ten bits, yielding 82 megabytes per data image. The shuttle elevation data form a 200Mbyte raw data set that needs to be filtered each time to yield a 6000 x 6000 point image. The seismic database is 1–2 gigabytes, taking tens of hours on a VAX to reduce to the pertinent information needed for a single image.

Once the three databases have been filtered, it takes yet more computer power to combine them to yield a rendered image. On a VAX, for example, it takes 14–17 minutes per frame for rendering. A minimal animation would be 1400 frames, requiring over 16 days of computing time.

The strategy to solve this problem is to break the problem down. Rendering of the data is performed on JPL's CRAY X-MP/18. The actual data filtering is done at Caltech, SDSC, and LANL. Figure 1 shows the extent of the data filtering. Even with the processing done at remote sites, there is still, if you have only one frame per minute, roughly 800 Mbps of data flow to the JPL.

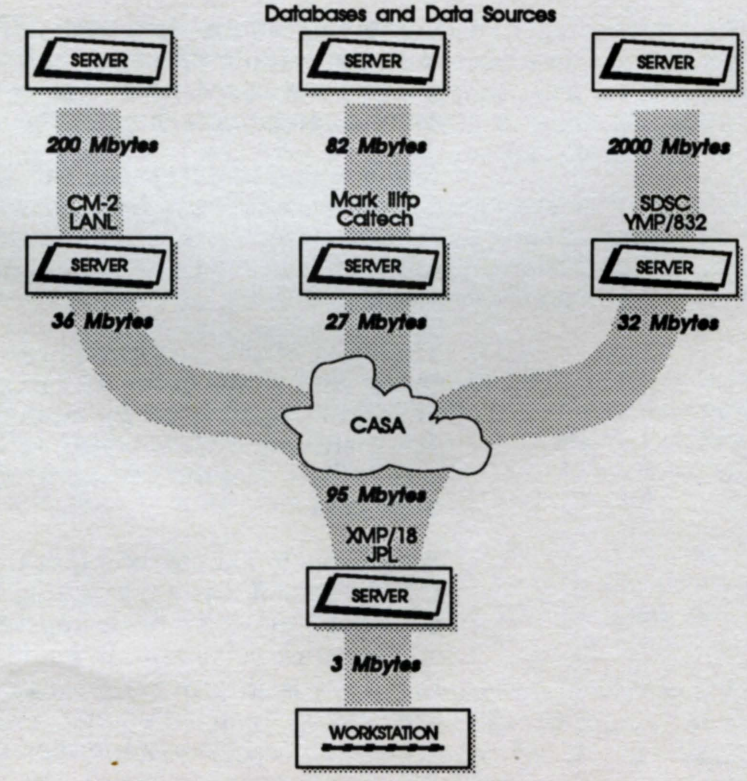

Figure 1: Dataflow in the CASA network

If you wanted to do real animation, it would take a minimum of 30 frames per second, yielding a data rate of 23.5 Gbps. The amount of processing power to do this animation would be 63 Gigaflops (the four machines involved in the application deliver around two Gigaflops of processing power).

The CASA network

The CASA network is a wide-area network. Each of the participating sites has a high-speed LAN, based on HIPPI. The host computers are all connected to the HIPPI switch. A HIPPI-SONET gateway is connected to the HIPPI switch.

The HIPPI-SONET gateway hooks up to long-haul optical fiber running the SONET protocols at STS-24 speeds (1.244 Gbps). Notice that SONET is being used directly instead of using an intervening ATM-based data link.

Linking HIPPI to SONET poses at least two problems. First, there is a difference in speed, with HIPPI running at 800 Mbps. Aside from rate adaptation, there is the more crucial problem of hiding latency. HIPPI won't let a source send data unless it has a ready signal. With a host required to store 64 ready HIPPI signals, and the propagation speed of HIPPI, we have a maximum HIPPI limit of 64 kilometers. CASA, however, needs 1500 to 2000 kilometers to function.

For HIPPI switches, CASA uses a switch developed at LANL in collaboration with DEC. The switch is a physical cross-bar switch; the switch actually moves to make the connection. This is a very fast physical switch, however, allowing a connection to be made in five microseconds if there is no contention.

The fiber for CASA is furnished by three telephone companies: MCI for the long-haul portion and the relevant Bell Operating Companies for the local loops. Built on top of this substrate is, to begin with, straight TCP/IP. If TCP proves inadequate as a transport layer, other candidates such as VMTP and NETBLT might be tried.

None of this, however, will be seen by the application programmer. The programmer would see, at the very lowest level, the UNIX sockets interface to TCP. Most programmers would work at even higher levels, using a collection of library routines such as *Express*. Express was designed for embedding information in **C** or FORTRAN programs that run on massively parallel processors. Express handles the questions of moving data and messages around and includes a symbolic parallel debugger and performance monitor for testing applications.

Serious?

Is the project serious? In addition to tremendous manpower, it is interesting to look at how much CPU time has been allocated on the big machines:

- SDSC has allocated 1900 CPU hours on the Y-MP 8/864.
- LANL has allocated 1100 hours on their Cray computer and 1100 hours on the CM-2.

The Cray computer CPU hours are supplemented by numerous other computers, not to mention a 1.2-Gbps, 1300-km fiber line.

CARL MALAMUD is the author of *DEC Networks and Architectures* (McGraw-Hill, 1989), *INGRES* (Van Nostrand Reinhold, 1989), *Analyzing Novell Networks* (Van Nostrand Reinhold, 1990), and *Analyzing DECnet/OSI Phase V* (Van Nostrand Reinhold, 1991). He has just finished writing *STACKS*, the official book of INTEROP 91, from which this article is adapted. The official caffeine-free diet soda of INTEROP 91 has not yet been chosen. Carl can be reached as `carl@malamud.com`.